Titre : Gestion de la performance et de la qualité de la détection parallèle des anomalies à large échelle

(English version below)

Mots-clés : Détection des anomalies, Données Complexes, Données distribuées, Passage à l'échelle, Performances, Parallélisation.

Encadrants : Dr. Seif-Eddine BENKABOU (seif.eddine.benkabou@univ-poitiers.fr), Dr. Amin MESMOUDI (amin.mesmoudi@univ-poitiers.fr) et Pr. Allel HADJALI (allel.hadjali@ensma.fr)

1. Contexte

La détection automatique d'anomalies joue un rôle critique dans une variété de domaines, notamment la cybersécurité, la maintenance prédictive et la surveillance de systèmes complexes. Les progrès récents dans le domaine de l'apprentissage automatique ont ouvert de nouvelles perspectives pour le développement de méthodes prometteuses de détection d'anomalies, en particulier lorsqu'il s'agit de données complexes telles que les graphes [1,5,6,7,8] et les séries temporelles [2,3,4].

Dans ce contexte, la gestion efficace de gros volumes de données est devenue cruciale pour la détection d'anomalies à grande échelle. Les systèmes modernes génèrent une quantité massive de données en temps réel [6], ce qui rend impératif d'adapter les méthodes de détection d'anomalies pour traiter ces flux de données de manière efficace. Cela nécessite non seulement une préparation de données efficace pour nettoyer, intégrer et étiqueter les données, mais aussi une parallélisation habile de l'exécution des algorithmes de détection d'anomalies. En tirant parti de la puissance du calcul distribué et des infrastructures de traitement de données à grande échelle, nous pourrons améliorer la réactivité et l'évolutivité de nos approches de détection d'anomalies, ce qui est essentiel pour répondre aux besoins des applications modernes. Par conséquent, la conception des algorithmes de détection d'anomalie devrait, en plus de la prise en compte des questions liées à la qualité de la détection, se pencher sur des questions telles que : comment stocker, organiser et indexer les données complexes ? Comment combiner l'indexation et la gestion de la mémoire pour des jeux de données extrêmement volumineuses, distribuées et multidimensionnelles ?

2. Défis scientifiques : Détection des anomalies à large échelle

Les techniques de détection d'anomalies doivent évoluer pour prendre en considération les environnements modernes de déploiement et ainsi faire face aux nouveaux défis engendrés par les données massives. Dans cette thèse, les contributions scientifiques attendues sont principalement liées à :

1) l'identification des goulots d'étranglement entravant les techniques de détection d'anomalies actuelles pour leur permettre de passer à l'échelle, et
2) le développement de nouvelles techniques de détection des anomalies qui prennent en charge la parallélisation massive des traitements sur de vastes volumes de données.

3. Profil recherché

Le candidat recherché devrait :

1. Être titulaire d'un diplôme de niveau Bac +5 en informatique ou en mathématiques appliquées, avec un intérêt pour la recherche.
2. Posséder une expertise en Machine Learning et en gestion de données à large échelle.
3. Avoir des compétences analytiques avancées et une capacité à résoudre des problèmes complexes.
4. Posséder une aptitude à communiquer à l'oral et à l'écrit en français et en anglais.

4. Candidature

Pour toute candidature, merci de vous adresser aux encadrants de thèse avant le **11 mars 2024** : Dr. Seif-Eddine BENKABOU (seif.eddine.benkabou@univ-poitiers.fr), Dr. Amin MESMOUDI (amin.mesmoudi@univ-poitiers.fr) et Pr. Allel HADJALI (allel.hadjali@ensma.fr).

Documents à fournir :

- Curriculum Vitae et lettre de motivation
- Notes de Master et/ou du diplôme d'ingénieur
- Tout document jugé pertinent par le candidat pouvant enrichir le dossier de candidature.

5. Références

[1] Xiaoxiao Ma, Jia Wu, Shan Xue, Jian Yang, Chuan Zhou, Quan Z. Sheng, Hui Xiong, Leman Akoglu. A Comprehensive Survey on Graph Anomaly Detection With Deep Learning. IEEE Trans. Knowl. Data Eng. 35(12): 12012-12038 (2023)

[2] Ane Blázquez-García, Angel Conde, Usue Mori, José Antonio Lozano. A Review on Outlier/Anomaly Detection in Time Series Data. ACM Comput. Surv. 54(3): 56:1-56:33 (2022)

[3] Seif-Eddine Benkabou, Khalid Benabdeslem, Vivien Kraus, Kilian Bourhis, Bruno Canitia. Local Anomaly Detection for Multivariate Time Series by Temporal Dependency Based on Poisson Model. IEEE Trans. Neural Networks Learn. Syst. 33(11): 6701-6711 (2022)

[4] Seif-Eddine Benkabou, Khalid Benabdeslem, Bruno Caniti. Unsupervised outlier detection for time series by entropy and dynamic time warping. Knowl. Inf. Syst. 54(2): 463-486 (2018)

[5] Jundong Li, Harsh Dani, Xia Hu, Huan Liu. Radar. Residual Analysis for Anomaly Detection in Attributed Networks. IJCAI 2017: 2152-2158

[6] Amin Mesmoudi, Mohand-Saïd Hacid, Farouk Toumani. Benchmarking SQL on MapReduce systems using large astronomy databases. Distributed Parallel Databases 34(3): 347-378 (2016)

[7] Abdallah Khelil, Amin Mesmoudi, Jorge Galicia, Ladjel Bellatreche, Mohand-Saïd Hacid, Emmanuel Coquery. Combining Graph Exploration and Fragmentation for Scalable RDF Query Processing. Inf. Syst. Frontiers 23(1): 165-183 (2021)

[8] Ishaq Zouaghi, Amin Mesmoudi, Jorge Galicia, Ladjel Bellatreche, Taoufik Aguili. GoFast. Graph-based optimization for efficient and scalable query evaluation. Inf. Syst. 99: 101738 (2021)

[9] Houssameddine Yousfi, Amin Mesmoudi, Allel Hadjali, Houcine Matallah, Seif-Eddine Benkabou. SRDF_QDAG: An efficient end-to-end RDF data management when graph exploration meets spatial processing. Comput. Sci. Inf. Syst. 20(4): 1311-1341 (2023)

Title: Managing performance and quality in Large-Scale Parallel Anomaly Detection

Keywords: Anomaly Detection, Complex Data, Distributed Data, Scalability, Performance, Parallelization.

Supervisors: Seif-Eddine BENKABOU (seif.eddine.benkabou@univ-poitiers.fr), Amin MESMOUDI (amin.mesmoudi@univ-poitiers.fr), and Allel HADJALI (allel.hadjali@ensma.fr)

1. Context

Anomaly detection is critical in a variety of fields, including cybersecurity, predictive maintenance, and complex system monitoring. Recent advances in machine learning have opened up new possibilities for developing promising anomaly detection methods, particularly with complex data such as graphs [1,5,6,7,8] and time series [2,3,4].

In this context, efficiently managing large data volumes is crucial for large-scale anomaly detection. Modern systems generate massive amounts of real-time data [6], necessitating the adaptation of anomaly detection methods to efficiently process these data streams. This requires not only effective data preparation for cleaning, integrating, and labeling but also the skillful parallelization of anomaly detection algorithm execution. Leveraging distributed computing power and large-scale data processing infrastructures will enhance the responsiveness and scalability of the anomaly detection approaches, crucial for modern applications. Thus, the design of anomaly detection algorithms should, in addition to addressing detection quality issues, focus on how to store, organize, and index complex data and how to combine indexing and memory management for large, distributed, and multidimensional data sets.

2. Scientific Challenges: Large-Scale Anomaly Detection

Anomaly detection techniques must evolve to address the challenges posed by massive data in modern deployment environments. The expected scientific contributions in this thesis primarily involve:

1. Identifying bottlenecks that hinder current anomaly detection techniques from scaling up, and
2. Developing new anomaly detection techniques that support the massive parallelization of processing over vast volumes of data.

3. Candidate Profile

The desired candidate should:

- Hold a Bachelor's degree (Bac +5) in computer science or applied mathematics, with an interest in research.

- Possess expertise in Machine Learning and large-scale data management.
- Have advanced analytical skills and the ability to solve complex problems.
- Possess the ability to communicate orally and in writing in both French and English.

4. Application

For any application, please contact the thesis supervisors before March 11, 2024: Dr. Seif-Eddine BENKABOU (seif.eddine.benkabou@univ-poitiers.fr), Dr. Amin MESMOUDI (amin.mesmoudi@univ-poitiers.fr), and Prof. Allel HADJALI (allel.hadjali@ensma.fr).

Documents to be provided:

- Curriculum Vitae and letter of motivation
- Master's and/or engineering degree transcripts
- Any document considered relevant by the candidate that could enrich the application dossier.

5. References

[1] Xiaoxiao Ma, Jia Wu, Shan Xue, Jian Yang, Chuan Zhou, Quan Z. Sheng, Hui Xiong, Leman Akoglu. A Comprehensive Survey on Graph Anomaly Detection With Deep Learning. IEEE Trans. Knowl. Data Eng. 35(12): 12012-12038 (2023)

[2] Ane Blázquez-García, Angel Conde, Usue Mori, José Antonio Lozano. A Review on Outlier/Anomaly Detection in Time Series Data. ACM Comput. Surv. 54(3): 56:1-56:33 (2022)

[3] Seif-Eddine Benkabou, Khalid Benabdeslem, Vivien Kraus, Kilian Bourhis, Bruno Canitia. Local Anomaly Detection for Multivariate Time Series by Temporal Dependency Based on Poisson Model. IEEE Trans. Neural Networks Learn. Syst. 33(11): 6701-6711 (2022)

[4] Seif-Eddine Benkabou, Khalid Benabdeslem, Bruno Caniti. Unsupervised outlier detection for time series by entropy and dynamic time warping. Knowl. Inf. Syst. 54(2): 463-486 (2018)

[5] Jundong Li, Harsh Dani, Xia Hu, Huan Liu. Radar. Residual Analysis for Anomaly Detection in Attributed Networks. IJCAI 2017: 2152-2158

[6] Amin Mesmoudi, Mohand-Saïd Hacid, Farouk Toumani. Benchmarking SQL on MapReduce systems using large astronomy databases. Distributed Parallel Databases 34(3): 347-378 (2016)

[7] Abdallah Khelil, Amin Mesmoudi, Jorge Galicia, Ladjel Bellatreche, Mohand-Saïd Hacid, Emmanuel Coquery. Combining Graph Exploration and Fragmentation for Scalable RDF Query Processing. Inf. Syst. Frontiers 23(1): 165-183 (2021)

[8] Ishaq Zouaghi, Amin Mesmoudi, Jorge Galicia, Ladjel Bellatreche, Taoufik Aguili.

GoFast. Graph-based optimization for efficient and scalable query evaluation. Inf. Syst. 99: 101738 (2021)

[9] Houssameddine Yousfi, Amin Mesmoudi, Allel Hadjali, Houcine Matallah, Seif-Eddine Benkabou. SRDF_QDAG: An efficient end-to-end RDF data management when graph exploration meets spatial processing. Comput. Sci. Inf. Syst. 20(4): 1311-1341 (2023)